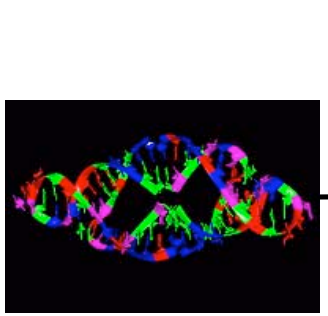
How we learned to cope with molecular biology data

Annalisa Barla

molecular biology

- a typical scenario is n<<d
- number of samples cannot always be increased (rare diseases and expensive technology)
- (mostly) high-throughput data
  - new technologies (DNA microarrays, SNP, CGH etc.)
  - possibility to measure the whole genome
  - most of the times the data are noisy (getting better any day now..)



biological samples

microarray gene expression

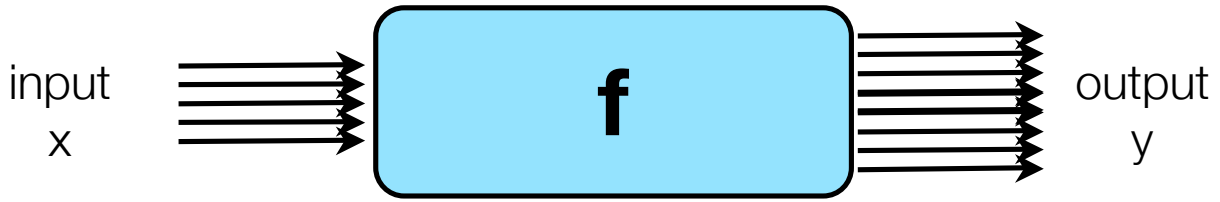
computational methods

202218_at	100
1556128_at	100
210788_at	88
…	…

selected genes

learning from example paradigm

the GOAL is not to memorize but to GENERALIZE, e.g. predict



given a set of **examples:**

$$\{(x_1,y_1),\ (x_2,y_2),\ldots,\ (x_n,y_n)\}$$

find a function:

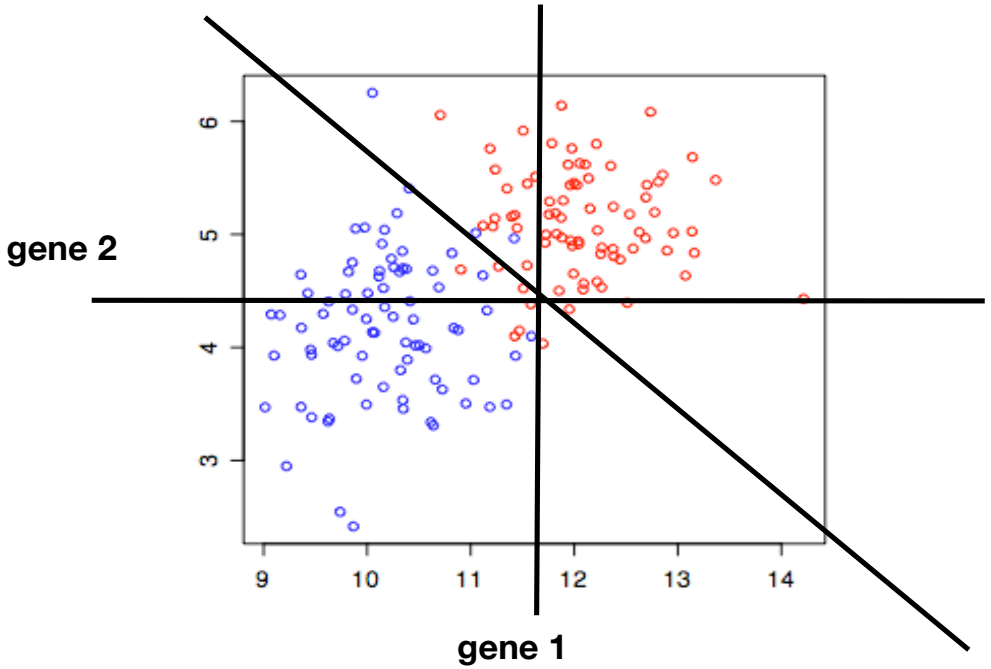
$$f(x)\sim y$$

such that **f** is a **good predictor on new data** as well as on the given dataset

and possibly identify the most discriminating variables (genes)
--> gene signature

why going multivariate?

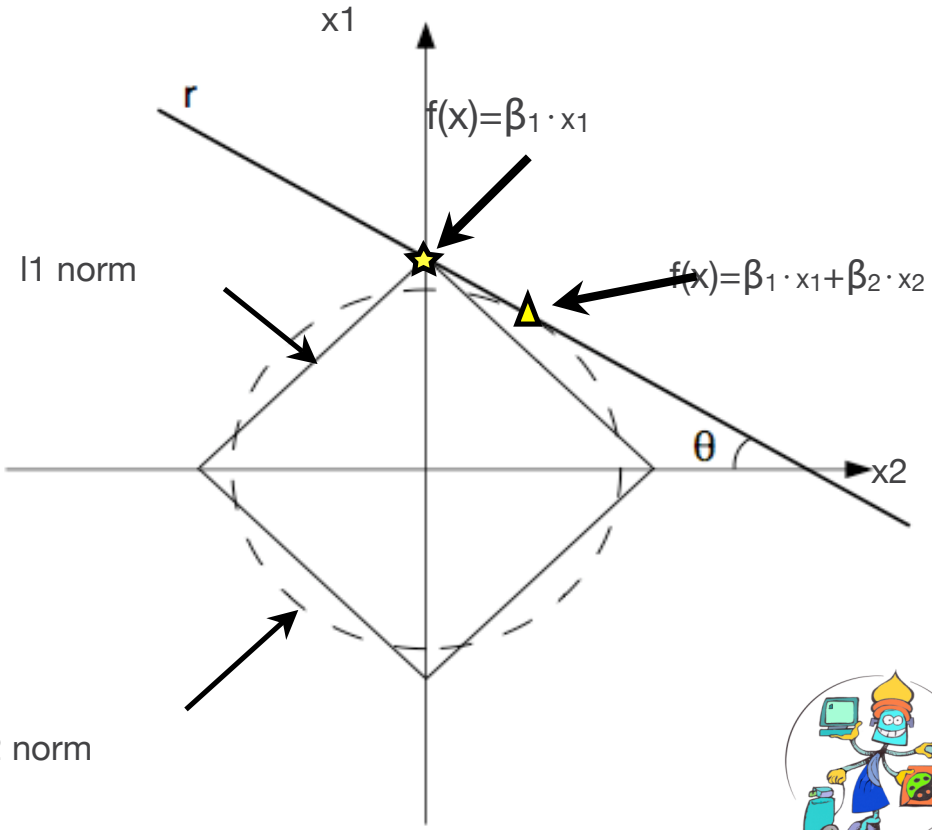
- search for **DIFFERENTIALLY EXPRESSED GENES** is not always sufficient!univariate approaches may not be flexible enough...



variable selection method

- Empirical Risk minimization combined with a mixed penalty:
  - l1 term enforcing sparsity
  - l2 term preserving correlation

$$\phi_{\tau,\mu} = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \tau||\boldsymbol{\beta}||_1 + \mu||\boldsymbol{\beta}||_2^2$$



Zou, H, Hastie, T.
Regularization and variable selection via the elastic net.
Journal of the Royal Statistical Society, 2005.

De Mol, C. Devito, E., Rosasco, L.***Elastic-net regularization in learning theory***Journal of Complexity, 2009

variable selection method

- Empirical Risk minimization combined with a mixed penalty:
  - l1 term enforcing sparsity
  - l2 term preserving correlation

regularization parameter

correlation parameter

$$\phi_{\tau,\mu} = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \tau||\boldsymbol{\beta}||_1 + \mu||\boldsymbol{\beta}||_2^2$$

- **Consistency guaranteed** (the more samples available the better the estimator)
- **Not univariate: takes into account behavior of many genes at once.**

variable selection method

$$\phi_{\tau,\mu} = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \tau||\boldsymbol{\beta}||_1 + \mu||\boldsymbol{\beta}||_2^2$$

- **output:** One-parameter family of nested lists with equivalent prediction ability and increasing correlation among genes.
- $\mu \to 0$: minimal list of prototype genes
- $\mu_1 < \mu_2 < \mu_3 < \ldots$: longer lists including correlated genes

since we have a **correlation parameter** we can tune and vary the list length

Christine De Mol, Sofia Mosci, Magali Traskine, Alessandro Verri.
A Regularized Method for Selecting Nested Groups of Relevant Genes from Microarray Data.
Journal of Computational Biology, 2009.

two stage approach (De Mol, Mosci, Traskine, Verri 2009)

variable selection step (l1l2):

$$||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \tau||\boldsymbol{\beta}||_1 + \mu||\boldsymbol{\beta}||_2^2$$

correlation parameter

classification step (rls):

$$\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

Given
-(X,Y) training set, and (X^{test}, Y^{test}) test set
-$\{(X_1,Y_1),\ldots,(X_k,Y_k)\}$ partition of (X, Y)
-$\mu_0 < \mu_1 < \ldots < \mu_{m-1}$

Stage I

-let $\mu = \mu_0$, $(\tau_t, \lambda_l)_{t\in\mathcal{T},\ l\in\mathcal{L}}$ a grid in parameter space
-for $t \in \mathcal{T}$ and $l \in \mathcal{L}$
  for $i = 1$ to k let
    $X_i^{tr} := X_1, \ldots X_{i-1}, X_{i+1}, \ldots, X_k$
    $Y_i^{tr} := Y_1, \ldots Y_{i-1}, Y_{i+1}, \ldots, Y_k$
    $\beta(t, l, i) :=$ classifier built on (X_i^{tr}, Y_i^{tr}) for $\tau = \tau_t$, $\mu = \mu_0$, and $\lambda = \lambda_l$
    $Err(t, l, i) :=$ error made by $\beta(t, l, i)$ on (X_i,Y_i)
  end
  $\overline{Err}(t, l) := \frac{1}{k}\sum_{i=1}^{k} Err(t, l, i)$
end

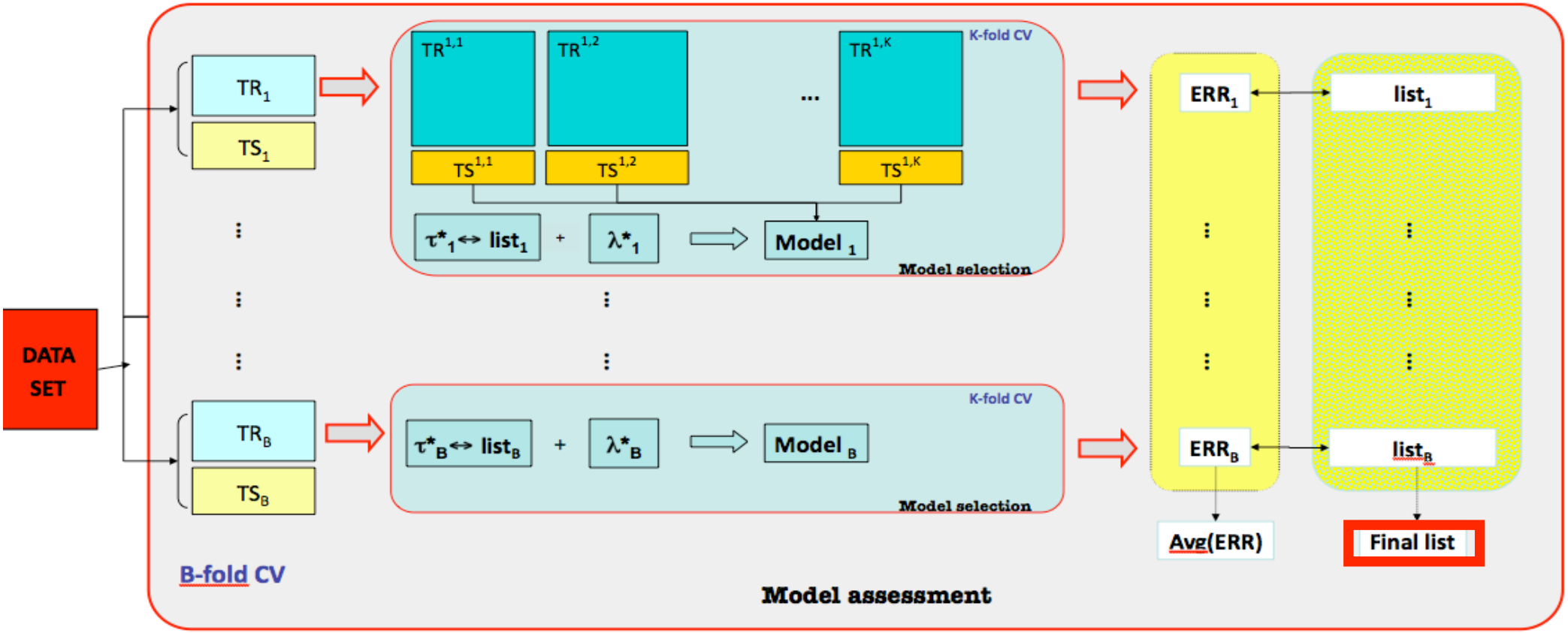
Stage II

-find $(\tau_{opt}, \lambda_{opt})$ minimizing $\overline{Err}(t, l)$
-for $i = 0$ to $m - 1$ let
    $\beta_{\mu}^* :=$ classifier built on (X,Y) for $\tau = \tau_{opt}$, $\mu = \mu_i$, and $\lambda = \lambda_{opt}$
    $Err_i^{test} :=$ error made by $\beta_{\mu_i}^*$ on (X^{test}, Y^{test})
end

for each **μ**
we have to **choose λ** and **τ**

statistical learning framework

$\lambda \rightarrow (\lambda_1, ..., \lambda_A)$
$\tau \rightarrow (\tau_1, ..., \tau_B)$



the optimal pair (λ^*, τ^*) is one of the $A \cdot B$ possible pairs $(\lambda, \tau)_{ij}$

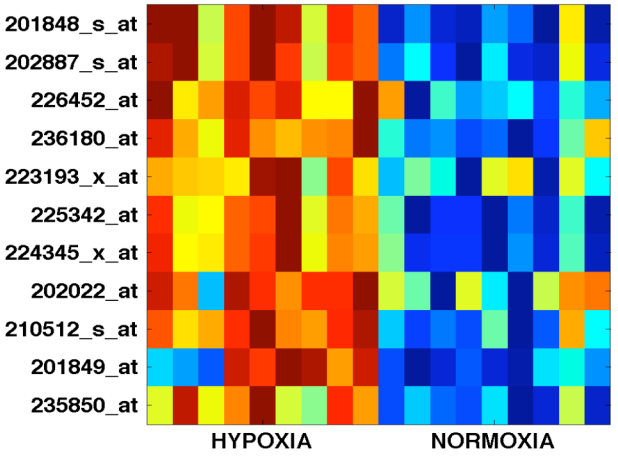
computational time in the LOO case (for one task):

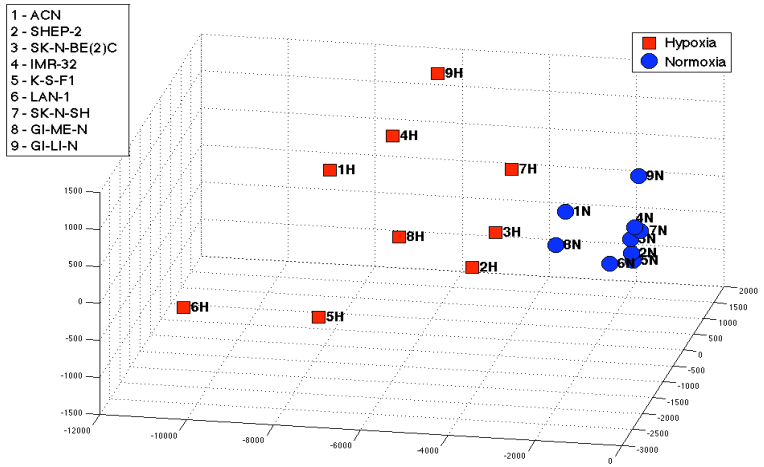
$time_{1\text{-optim}} = (2.5s \div 25s)$ depending on the correlation parameter

Total Time = $A \cdot B \cdot N.samples \cdot time_{1\text{-optim.}} \sim 20 \cdot 20 \cdot 30 \cdot time_{1\text{-optim}} \sim 2 \cdot 10^4 s \div 2 \cdot 10^5$

Identifying the Hypoxia Signature of Neuroblastoma via Regularization

- partner: IGG Molecular Biology lab
- Dataset: 9 neuroblastoma (NB) cell lines cultured under normoxic (normal O_2) and hypoxic conditions (low O_2).
- Technology: Affymetrix GeneChip U133 plus 2.0. (~54000 variables)
- t-test: no genes selected!
- I1I2 protocol: 11 genes for the minimal list (frequency>30%)





Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Varesio L
BMC Genomics 2009, 10:474 (15 October 2009)

a step forward: prior (GO)

- Specific prior knowledge can be used to better understand the biological phenomenon under study
- Possible sources:
  - Digital online data libraries
  - Textbook knowledge
  - MDs / experts

We selected subset of variables from the Gene Ontology and performed the variable selection on those sub-matrices

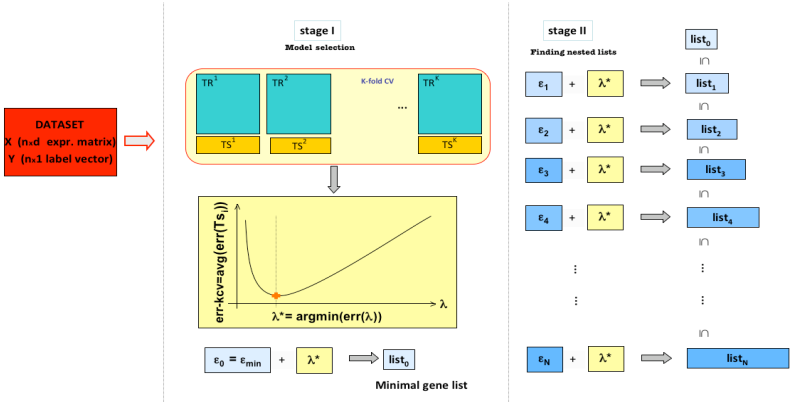
- hypoxia related groups
- MYCN related groups
- neuroblastoma related groups

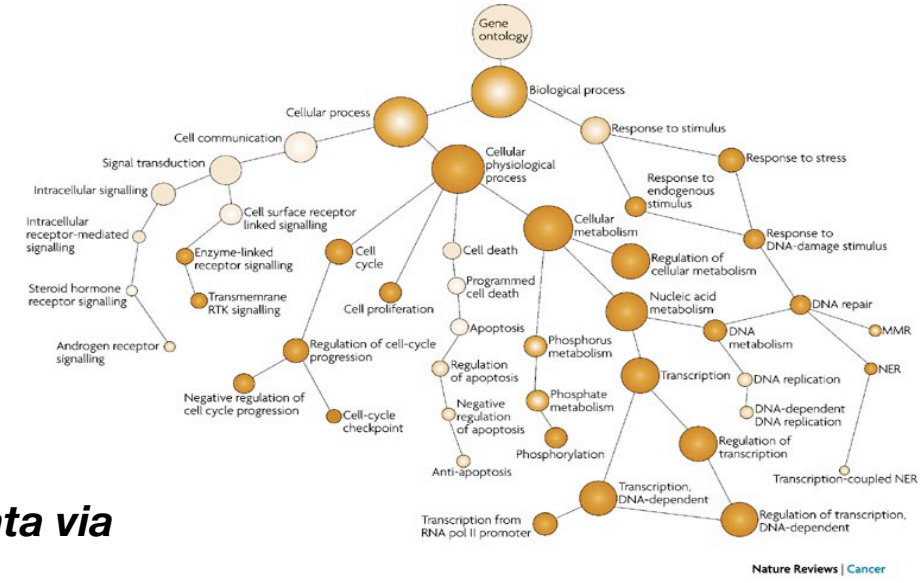
Fardin, P Cornero, A Barla, A Mosci, S Acquaviva, M Rosasco, L Gambini, C Verri, A and Varesio, L
Journal of Biomedicine and Biotechnology 2010 (to appear)

another step forward: Function-based analysis of microarray data via l1-l2 regularization

- combine the selection protocol with the GO structure automatically

- provide a way to easily interpret the output of feature selection protocol

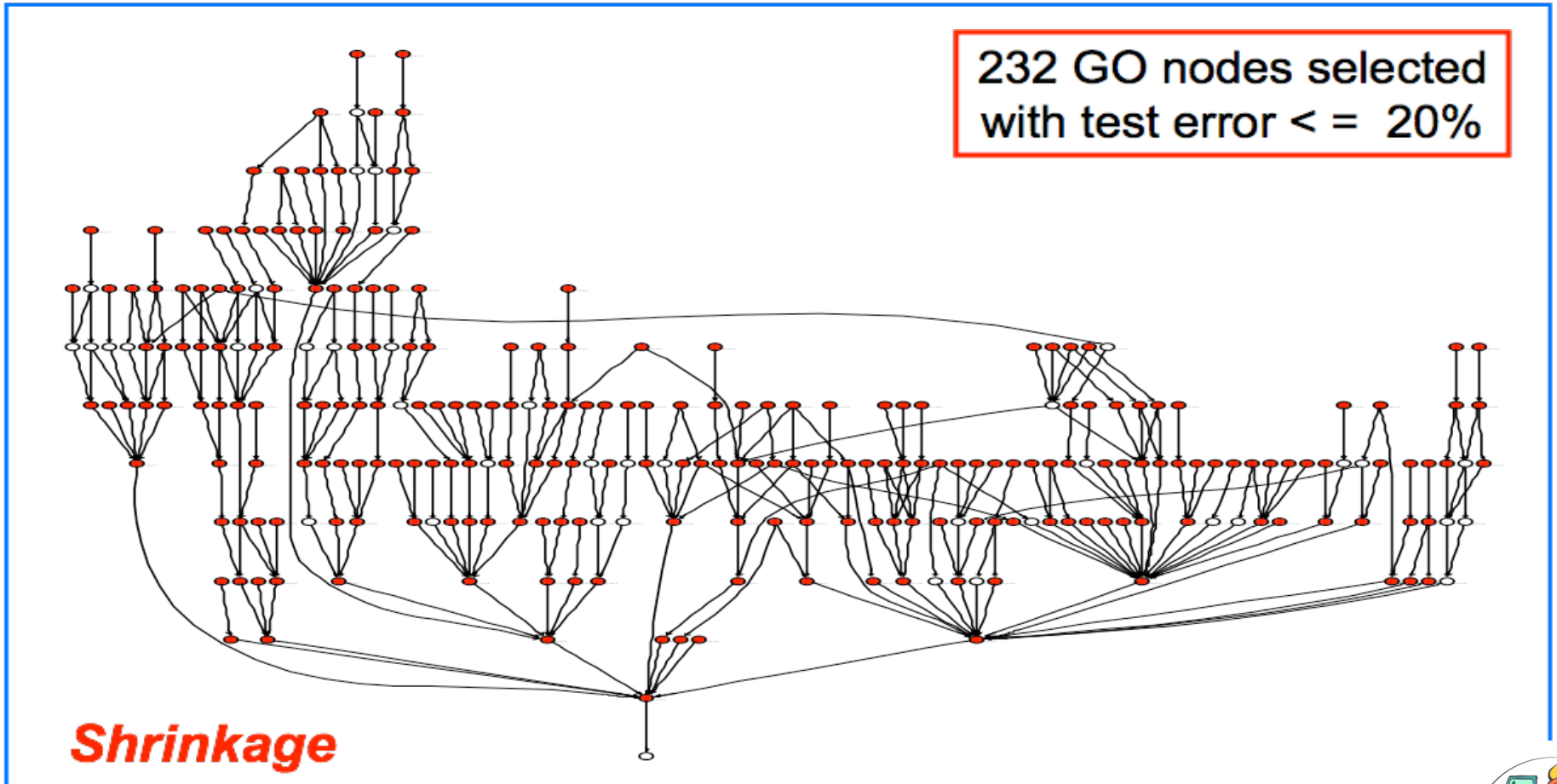




joint work with University of Padova

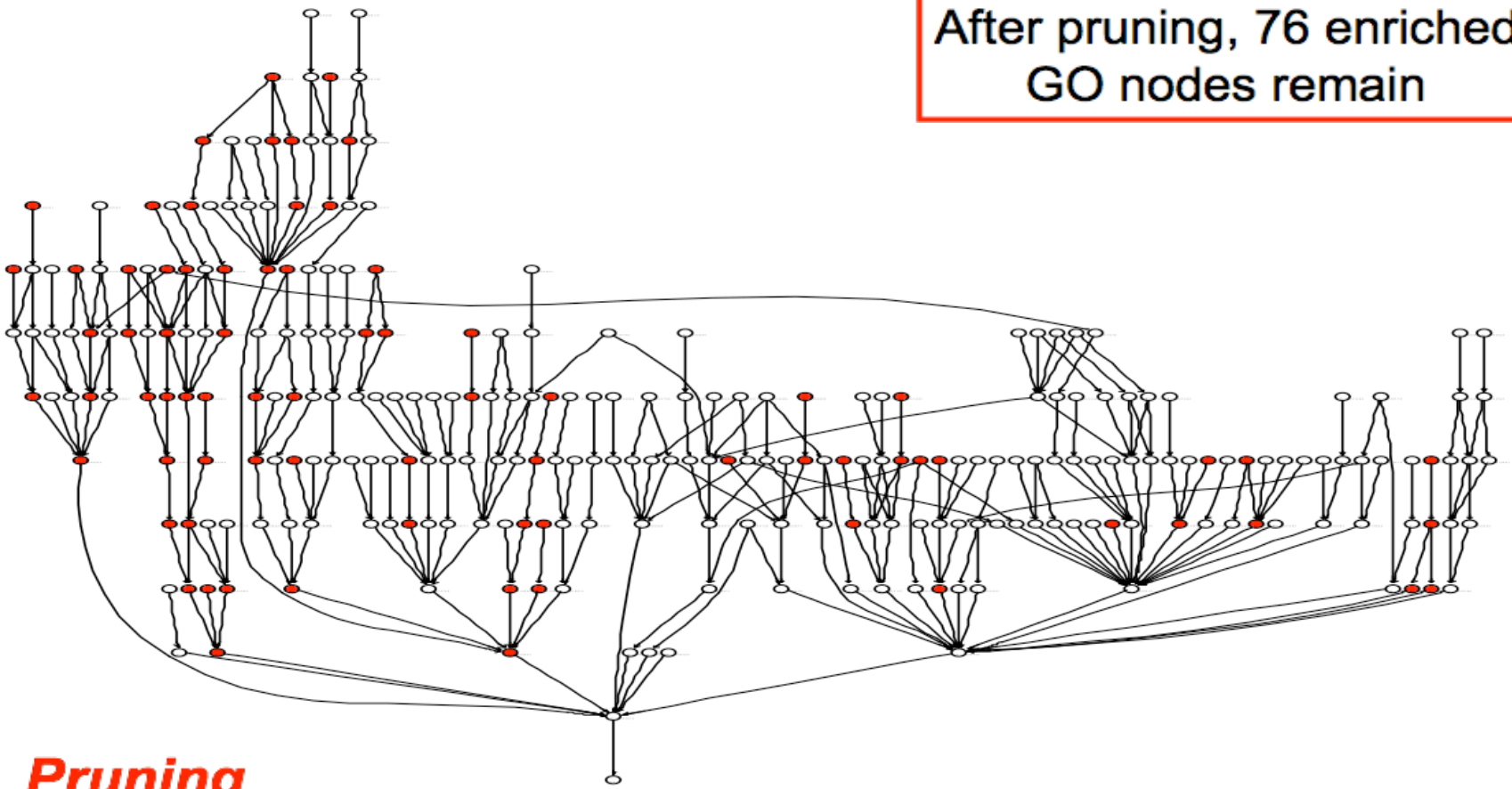
Function-based analysis of microarray data via l1-l2 regularization
poster @ ECCB'09

case study: breast cancer data (GSE7390)



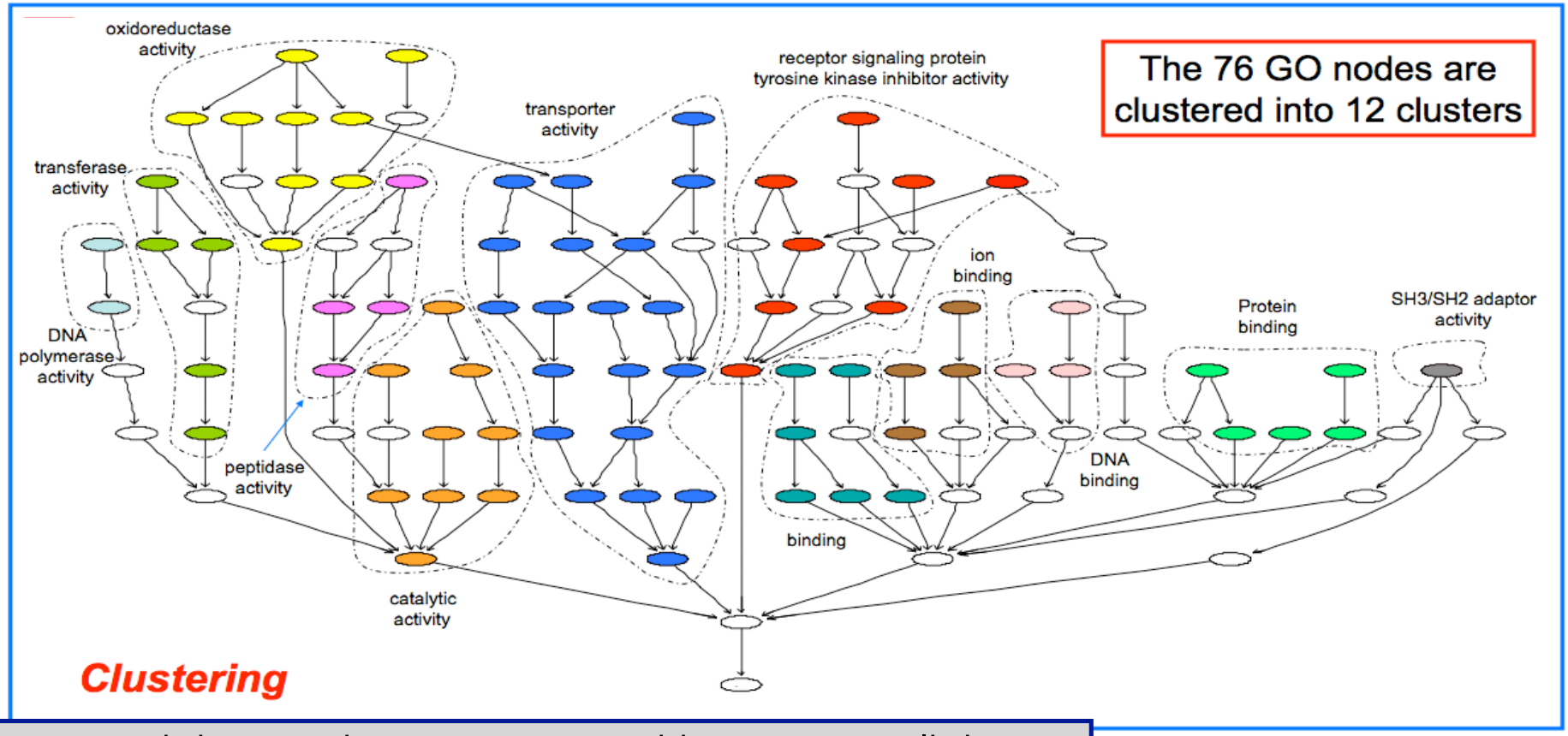
Selected GO nodes after l1l2 feature selection step

case study: breast cancer data



Selected GO nodes after enrichment test
(selected genes/total genes in the node)

case study: breast cancer data



The remaining nodes are grouped by average linkage hierarchical clustering based on semantic similarity

what we learned

- beware of the selection bias
- go multivariate
- make use of the vast prior knowledge

- learn how to distribute the computation (grid/cloud/cluster)
- use open source software (in order to distribute on a cloud)

- some biology
- a common language with biologists and MDs

work in progress: methods to incorporate prior knowledge

- Kernel design
- Group lasso/Graph lasso (Jacob, L Obozinski, G Vert, JP)
- Semantic learning